сочетаниями должны исследоваться и другие критерии оценки текста для увеличения вероятности правильного установления авторства (каковы они – это предстоит выяснить в ходе дальнейших исследований).

Библиографические ссылки

1. Maurice van Keulen Matching Profiles from Social Network Sites, 2009 [Электронный ресурс]. URL:http://wwwhome.cs.utwente.nl/~keulen/wordpress/2009/10/matching-profiles-from-social-network-sites/

2. *Андреев Н. Д.* Статистико-комбинаторные методы в теоретическом и прикладном языковедение. М. : Наука, 1967. 402 с.

3. *Валгина Н. С.* Теория текста : учеб. пособие. М. : Логос, 2003. 280 с.

4. *Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф.* Справочник по теории вероятности и математической статистике. М. : Наука, 1985. 640 с.

ФОРМИРОВАНИЕ БАЗЫ РЕШАЮЩИХ ПРАВИЛ СИСТЕМЫ ОБНАРУЖЕНИЯ АТАК С ПОМОЩЬЮ ГЕНЕТИЧЕСКОГО АЛГОРИТМА

А. О. Власов
(Екатеринбург, УрФУ, ale_vlas@mail.ru)

Формирование базы решающих правил для системы обнаружения атак является одной из главных задач при защите от компьютерных атак. Качественно построенная система правил позволяет выявлять и предотвращать опасные или потенциально опасные воздействия на автоматизированные информационные системы.

Для пользователей и администраторов готовых к использованию «коробочных» систем обнаружения атак формирование и разработка правил распознавания не являются важной задачей: производители зачастую не включают в функционал подобных систем использование пользовательских правил, а имеющиеся базы регулярно обновляются с серверов разработчиков. Задача формирования базы правил в первую очередь интересна производителям сис-

тем обнаружения атак для своевременного построения защитных механизмов от регулярно появляющихся в сети новых видов атак. От скорости реагирования на вновь появляющиеся атаки и выпуска новых баз правил напрямую зависит репутация производителя и, соответственно, прибыль компании. Также решение этой задачи востребовано там, где необходимо защищать ценную и важную информацию силами собственных служб информационной безопасности, где ожидание обновлений баз производителей является критичным или компьютерные атаки являются специфичными и таргетированными.

Ясно, что ключевыми характеристиками успешности решающих правил является их качество распознавания (максимальное количество корректно распознанных атак, минимальное число ложных тревог) и время от разработки до внедрения в систему. Экспертный метод разработки правил является наиболее очевидным и популярным способом решения задачи, однако он имеет существенные недостатки: ввиду наращиваемых вычислительных мощностей и квалификации злоумышленников сложность компьютерных атак постоянно возрастает, и эксперту, даже группе экспертов, трудно качественно и своевременно обнаружить ключевые атакующие характеристики массивов трафика. Большую роль также играет опыт и квалификация экспертов. Логичным решением является построение математических моделей, алгоритмов поиска оптимальных правил и использование вычислительной мощности современных компьютеров. Тем самым можно свести к минимуму влияние экспертов, а также сократить время их работы. В данной работе будет представлена идея использования генетического алгоритма, который позволяет в сжатые сроки находить близкие к оптимальному решения задачи. Также будет предложена формальная математическая модель, постановка задачи в ее терминах, предложен алгоритм и результаты его работы.

Задача выделения ключевых характеристик трафика для систем обнаружения атак из набора всех возможных является задачей дискретной оптимизации. Алгоритм формирует базу решающих правил с помощью генетического алгоритма, который позволяет снизить размер пространства поиска методом полного перебора.

Математическая модель

Каждая характеристика трафика (IP-адрес источника, IP-адрес назначения, порт источника, порт назначения, длина пакета, количество сессий с узла и т. д.) является элементом множества значений этой характеристики.

Пакет или совокупность пакетов трафика целесообразно представить как упорядоченный набор элементов множеств значений входящих в него характеристик: $(x_1, ..., x_n) \in X_1 \times ... \times X_n$, где x_i – значение характеристики, а X_i – множество всех возможных значений для i-й характеристики. Так как все характеристики имеют различное множество значений (IP-адреса принимают значения из пространства пула адресов, TCP- и UDP-порты из множества {1, ..., 65535} и т. д.), целесообразно нормировать все множества так, чтобы каждое множество состояло из вещественных чисел от 0 до 1. То есть $X_i^{\text{норм}} = \left\{ \frac{x_i}{|X_i|}, i \in \overline{1, ..., n} \right\}$, где $|X_i|$ – мощность множества X_i.

Таким образом, каждый пакет представляется в виде точки пространства $(x_1, ..., x_n) \in q^n$, $q \in \square_{\{0...1\}}$.

Каждый пакет может являться атакующим или безопасным. Демаскирующим признаком атакующего пакета может являться какая-либо из его характеристик. Система правил алгоритма в терминах математической модели представляется в виде покрывающих множеств размерности от 1 (одна демаскирующая характеристика – например, большая длина пакета) до n (все характеристики являются ключевыми). Для покрытия характеристик использовались следующие множества:

$$L = \left\{ [0, x] \vee [x, y] \vee [y, 1] \right\}^k, \; x, y \in \{0 ... 1\}, \; k \in \{1 ... n\}. \qquad (1)$$

Качество правил оценивается следующим образом:

$$C = \frac{P}{|P|} - \frac{S}{|S|} - \frac{\dim L}{n}, \qquad (2)$$

где $\frac{P}{|P|}$ – доля корректно распознанных атакующих пакетов (количество распознанных атак на общее количество атакующих пакетов), $\frac{S}{|S|}$ – доля ложных тревог (количество ложных тревог на общее число «фоновых» пакетов), $\frac{\dim L}{n}$ – доля ключевых характеристик относительно общего числа характеристик («вес» правил – чем выше, тем сложнее и массивнее правило).

Сложность построения таких множеств – экспоненциальная, данная задача решаема методами дискретной оптимизации, таким образом, использование генетических алгоритмов для быстрого нахождения экстремумов является оправданным.

Генетические алгоритмы используют для работы эволюционные принципы наследственности, изменчивости и естественного отбора.

Алгоритм работает с популяцией особей, в которых закодировано возможное решение задачи (система правил). Каждая особь представляет собой некое правило, описанное в формуле (1).

В начале работы алгоритма случайным образом формируется начальная популяция. Далее посредством вероятностных механизмов выбирается множество пар для скрещивания особей. Обе особи, которые составляют родительскую пару, выбираются случайным образом из всей популяции, причем любая особь может стать родителем несколько раз. В результате скрещивания выбранных особей посредством применения генетического оператора кроссовера создается потомство, генетическая информация которого формируется в результате обмена генетической информацией между родительскими особями. В качестве оператора кроссовера экспериментальным путем был выбран арифметический двухточечный кроссовер:

A_k, B_k – k-е родительские гены;

a_k, b_k – k-е гены потомков;

$a_k = \lambda A_k + (1 - \lambda) \cdot B_k$;

$b_k = \lambda B_k + (1 - \lambda) \cdot A_k$, $\lambda \in (0, 1)$.

Созданные потомки формируют новую популяцию, часть потомков мутирует (используется генетический оператор мутации), что выражается в случайном изменении их генетической информации.

Этап, включающий последовательность «Оценивание популяции» – «Селекция» – «Скрещивание» – «Мутация», называется поколением. Эволюция популяции состоит из последовательности таких поколений.

Для того чтобы оценить качество закодированных решений, используется функция приспособленности, которая оценивает, насколько хорошо особь (или поколение) решает задачу. Постоянный контроль значений функции приспособленности поколений необходим для оценки прогресса выполнения поставленной задачи, а также анализа качества подобранных параметров генетического алгоритма. Также такая оценка может быть полезна для выбора «элитных» особей и сохранения их в популяции (мутация и скрещивание могут снизить приспособленность «элитных» особей; табл. 1).

Таблица 1

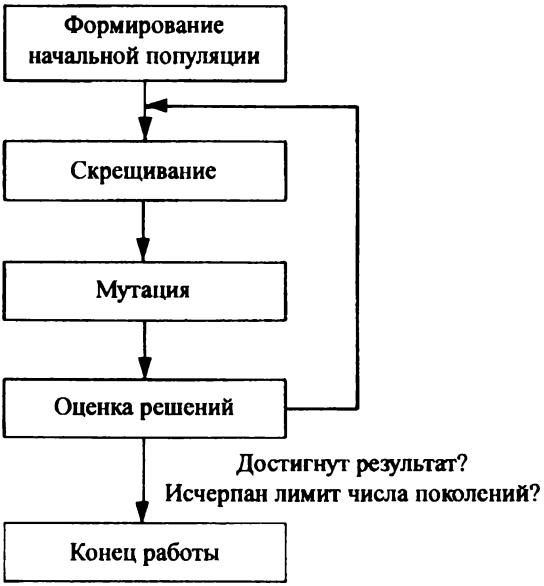
Параметры генетического алгоритма

№ п/п	Параметр	Значение
1	Количество особей в поколении	15
2	Вероятность мутации гена	0,075
3	Коэффициент кроссовера	0,5
4	Коэффициент «элитизма»	0
5	Коэффициент «новичков»	0,1
6	Выбор пары для кроссовера	Случайная пара
7	Оператор кроссовера	Арифметический кроссовер с $\lambda = 0{,}3$, двухточечный кроссовер

В качестве функции оценки приспособленности используется функция (2).

Концом работы алгоритма может служить достижение необходимого результата или исчерпывание заданного числа поколений.

Типичная схема работы генетического алгоритма представлена на следующем рисунке.



Тестирование работы генетического алгоритма

В табл. 2 представлены результаты работы генетического алгоритма на тестовых данных, содержащихся в наборе файлов. Каждая строка файла представляла собой аналог сетевого пакета, разбитого на характеристики, и была помечена кодом атаки (0 – в случае «фонового», нормального трафика). Массив трафика состоял из случайно сгенерированных 100 «пакетов», 20 % из которых являлись атакующими.

Все параметры выбраны эмпирическим путем, однако с помощью теоретических выкладок было значительно уменьшено пространство поиска наилучших параметров.

Таблица 2

Результаты работы генетического алгоритма

№ п/п	Количество характеристик	Поколений	% корректно распознанных атак	% ложных тревог
1	1	10 000	100	17,5
2	1	20 000	100	17,5
3	1	40 000	100	8,75
4	1	50 000	100	8,75
5	2	10 000	100	12,5
6	2	20 000	100	8,75
7	2	40 000	100	5
8	2	50 000	100	3,75
9	3	10 000	100	15
10	3	20 000	100	6,25
11	3	40 000	100	6,25
12	3	50 000	100	6,25
13	3	3 000 000	100	1,25
14	5	10 000	100	18,75
15	5	20 000	100	10
16	5	40 000	100	10
17	5	50 000	100	10
18	7	50 000	100	12,5
19	7	100 000	100	12,5

Из-за использования вероятностного алгоритма и чувствительности результатов к характеристикам исходного трафика, представленного в тестовых файлах, полученные результаты для одного массива данных не могут гарантировать результаты такого же уровня для другого массива (а также для эквивалентных массивов дан-

ных при повторном тестировании). Ухудшение результатов при увеличении числа поколений также объясняется вероятностной природой алгоритма.

Результат

Использование генетического алгоритма для формирования баз решающих правил позволяет за сравнительно короткое время сформировать достаточно точную систему правил, описывающую заданные виды атак. Сгенерировав с помощью «сплойтов» и «фонового» трафика обучающий массив, с помощью генетического алгоритма можно быстро построить систему правил, нейтрализующую данные атаки с минимальной долей ложных тревог и при этом уменьшить работу экспертов.

ОРГАНИЗАЦИЯ БЕЗОПАСНОГО ОБМЕНА ДАННЫМИ МЕЖДУ МИС И НЕЗАВИСИМОЙ ИНФОРМАЦИОННОЙ СИСТЕМОЙ ПО ПРОТОКОЛУ MEDML

А. М. Воробьев
(Тюмень, ТюмГУ, artandvor@gmail.com)

Важной социально значимой задачей является повышение доступности и качества медицинских услуг. Одним из направлений решения задачи является сокращение очередей в регистратурах, упрощение процесса записи на прием. Этому способствует применение медицинских информационных систем и организация возможности самозаписи пациентов через Интернет.

В медицинских учреждениях России используется множество медицинских информационных систем (МИС), разработанных различными компаниями, написанных на разных языках и имеющих разные структуры. Большинство медицинских систем не предусматривают возможность самозаписи через веб-сайт. Поэтому учреждения при предоставлении возможности организации записи через Ин-